# A Study of Several Classification Algorithms to Predict Students' Learning Performance

**Ms. SriUdaya Damuluri, George Mason University**

SriUdaya Damuluri is a graduate student in Data Analytics Engineering at George Mason University. Her research interests include big data, predictive analytics, and machine learning. She earned her bachelor's degree from Jawaharlal Nehru Technological University.

**Dr. Pouyan Ahmadi, George Mason University**

Pouyan Ahmadi is an Assistant Professor in the Department of Information Sciences and Technology.

His research interests include cooperative communications and networking, cross-layer design of wireless networks, relay deployment and selection in wireless networks. He has devised several relay-selection strategies in cooperative communication by combining routing and cooperative diversity with the consideration of a realistic channel model that can be applied to WLAN and LTE networks.

Dr. Ahmadi earned his Ph.D. in Electrical and Computer Engineering at George Mason University, M.S. in Architecture of Computer Systems at Iran University of Science and Technology, and B.S. in Computer Engineering at Azad University.

**Dr. Khondkar Islam, George Mason University**

Khondkar Islam is an Associate Professor and Associate Chair for Undergraduate Studies in the Department of Information Sciences and Technology at George Mason University (GMU).

His research interests include distributed and peer-to-peer systems, overlay and wireless networks, network security, engineering education, and distance education for instructor training and student learning.

Dr. Islam earned his Ph.D. and B.S. at George Mason University, and M.S. at American University.

# A Study of Several Classification Algorithms to Predict Students' Learning Performance


## Abstract

Identifying students who need better pedagogical support is an invaluable asset for any academic institution. The main objective of this study is to predict the students' performance and thereby maximize their learning productivity. We focus on the students' past academic performance to predict their future results. This is done by analyzing the various factors of course material and students' online behavior from the Learning Management System (LMS). We also analyze several predictors that contribute to the overall student performance from the data collected. To determine the efficient model that is more accurate and precise, we compare the performance of four well-known machine learning classification algorithms. The 2017 and 2018 academic year data collected consists of user patterns, navigational behavior and the students' daily activities from the LMS, Blackboard (Bb) Learn of the Undergraduate IT program within the Information Sciences and Technology (IST) Department at George Mason University (GMU). This comparison effort will help us confirm the most effective algorithm to identify students' who are at risk of failing a class so that academic advisors/instructors can offer better academic guidance and support.



Keywords: Classification algorithms, navigational behavior, performance prediction, Learning Management System.


## 1. Introduction

One of the major goals in any higher educational institution is to improve students' performance. A Learning Management System (LMS) can be used as a platform to assess students' performance. Several universities have been using LMS for the past few years which is a software application that helps to administer, track and deliver educational courses. This system has paved a way for educators to monitor students' online activities and learning behaviors to derive essential conclusions.

Several studies were conducted to examine the relationship between varying student performance and their online behavior using LMS. This work is an extension which further focuses on predicting the future performance of the students. This involves the investigation of students' data and employing appropriate machine learning algorithms to analyze the students' past work that can help in predicting the future performance of current students. We apply various classification algorithms to determine which algorithm predicts the performance with better accuracy. Support Vector Machine [1], Naïve Bayes [2], K-Nearest Neighbor [3], and Linear Discriminant Analysis [4] algorithms are used in this study. Using these algorithms, we create a training model that accurately predicts a target value by learning decision rule derived from prior data (training data).

In this study, we collect the data of students from an undergraduate networking course through Blackboard (Bb) Learn. Bb Learn is the LMS used at GMU for delivering course content which includes several modules like course materials, announcements, discussion boards, assignments, etc. Instructors use Bb's resources to upload the course materials and students access them for learning and to submit the assignments, labs, quizzes, and discussions. This study considers 300 students' data who enrolled in 11 sections of two identical networking courses offered in Volgenau School of Engineering at GMU. These courses are IT 341- Data Communications and Networking Principles and CYSE 230 - Computer Networking Principles, from Spring and Fall 2018 academic year.

## 2. Literature Review

There are several studies in the literature that have analyzed students' performance and predicted success rate based on their learning activities. One such study is done by Superby and Meskens [5] in which they demonstrate the application of neural networks, random forests, and decision trees in predicting the success rate of students. In this study, students were classified into three groups: low-risk, medium risk, and high risk. The most correlated factors were found to be attendance, previous academic experience and study skills.

The study by Hung and Zhang [6] analyzed various patterns in an online business course and found that the maximum logins were made on Tuesday which was the start date of weekly projects. This indicates that 26% of the students worked on their project right away. They also found that most logins were made during the first and last week of the course which gave some insights for instructors to better schedule the course content and deadlines.

Another study by Widyahastuti et al., [7] predicted the performance on the final exam using Linear Regression [8] and Multilayer Perceptron Network [9] in Weka which is a tool used for data mining tasks. This study focused on the data collected from online discussion forums and student attendance for prediction.

A machine learning framework was used in [10] to identify students who were at the risk of not graduating high school on time. Several training models were employed, and the performance of each model was compared using classification metrics like precision, recall, and area under curve (AUC) [11]. Another similar study [12] tried to develop an early warning system to predict students' online learning performance. The data was taken from an online course which concluded that CART (Classification and Regression Tree) is the best classifying algorithm to evaluate the learning performance. The study also concluded that time-dependent variables are crucial for achieving higher prediction accuracy.

The study by Macfadyen and Dawson [13] tried to explore the factors of student online activities that can predict academic achievement accurately. The study identified several variables that were correlated with the students' final grade. The total number of discussions, mail messages, and assessments completed were found to be the most correlated factors. Another study by Won You [14] found several indicators in which regular study was the strongest predictor followed by late submissions, discussion messages, frequency of course login and proof of reading the course information, played a significant role in predicting the online course achievement. Decision tree

method is used in classifying the students' data to predict the performance in [15] where attendance, class test, seminar, and assignment scores were considered as predictors for the student data.

This study is not limited to predict the performance alone. It focuses on calculating several algorithms' accuracy levels based on behavioral data of students accessing the course content in order to identify the most effective algorithm that can detect students who need help in improving their academics. This information can help educators and instructors to design the coursework in an efficient manner. We give attention to all the aspects of students' online behavior like number of hours spent on the course each day of the week, total logins to the course, number of hours spent and the number of times the student accessed course items.

## 3. Research Study and Analysis

The objective of this study is to compare several classification models and determine which algorithm works efficiently with regard to a number of evaluation metrics. The steps involved in the study are listed below:

A. Data collection
B. Data pre-processing
C. Feature selection
D. Training model process
E. Model evaluation

### A. Data collection

Data collection is one of the most important and time-consuming stages of this analysis. The quality and integrity of the data have to be maintained to get real and accurate predictions. The study began with the data collection of students' access behavior from Blackboard Learn. We made use of 11 sections from IT341 and CYSE230 courses offered in Spring and Fall 2018 semesters in Volgenau School of Engineering at GMU. Table 1 describes each attribute of the data extracted from Blackboard. The data was used in R Studio to perform data analysis and prediction.

**Table 1. Attributes collected from Blackboard**

| Attribute Name | Description |
|---|---|
| Sunday-Saturday | Time spent on a particular day |
| Total time | Time spent on the course |
| Total logins | Total logins |
| Total items | Items accessed |
| Ch1-Ch11 (Duration) | Time spent on each chapter |
| Ch1-Ch11 (Clicks) | No. of times each chapter is accessed |

| | |
|---|---|
| RS1-RS8 (Duration) | Time spent on Routing and Switching chapters |
| RS1-RS8 (Clicks) | No. of times Routing and Switching chapters are accessed |
| HA1 & HA2 (Duration) | Time spent on Homework Assignments |
| HA1 & HA2 (Clicks) | No. of times Homework Assignments are accessed |
| LS1-LS12 (Duration) | Total time spent on each lab session |
| LS1-LS12 (Clicks) | No. of times each lab session is accessed |
| Grade | Student course grade |

**B. Data Pre-processing**

Data pre-processing requires the dataset to be clean and consistent in order to achieve better performance for modeling. In order to make our data more consistent, we used Mean Imputation Method [16] which is done by replacing the missing values in the dataset with the mean of the variable. Figure 1 represents a portion of data after pre-processing in R Studio.

| | Sunday | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday | Total_time | Total_items | Total_logins | Ch1_Totaltime |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3.48 | 9.05 | 0.02 | 3.86 | 0.09 | 3.44 | 4.01 | 12.38 | 44 | 74 | 0.00 |
| 2 | 5.64 | 16.00 | 0.03 | 0.70 | 0.15 | 0.13 | 1.87 | 24.31 | 63 | 195 | 0.00 |
| 3 | 14.04 | 15.02 | 0.06 | 0.19 | 0.14 | 3.44 | 2.27 | 31.44 | 48 | 119 | 0.00 |
| 4 | 6.88 | 15.19 | 1.26 | 0.75 | 2.59 | 0.05 | 2.31 | 29.01 | 58 | 156 | 0.07 |
| 5 | 15.73 | 11.30 | 9.65 | 0.01 | 0.34 | 0.01 | 0.23 | 37.15 | 55 | 151 | 0.00 |
| 6 | 4.12 | 20.42 | 4.45 | 0.23 | 5.67 | 0.43 | 0.05 | 35.22 | 50 | 107 | 0.00 |
| 7 | 6.60 | 20.10 | 0.07 | 0.14 | 2.43 | 0.00 | 0.59 | 29.56 | 65 | 128 | 0.00 |
| 8 | 13.38 | 18.07 | 0.90 | 1.99 | 3.69 | 2.96 | 13.23 | 54.13 | 80 | 329 | 0.00 |
| 9 | 6.40 | 16.62 | 0.34 | 2.20 | 2.66 | 0.06 | 0.58 | 28.51 | 66 | 242 | 0.00 |
| 10 | 7.11 | 15.51 | 2.53 | 0.09 | 0.06 | 0.21 | 0.71 | 26.13 | 54 | 218 | 0.07 |
| 11 | 16.94 | 33.76 | 2.87 | 1.87 | 18.34 | 2.37 | 7.68 | 83.49 | 84 | 326 | 0.00 |
| 12 | 8.20 | 19.10 | 3.96 | 3.75 | 0.90 | 0.01 | 1.33 | 37.15 | 70 | 225 | 0.07 |

**Figure 1. Data after Pre-processing**

Classification is one of the predictive modeling techniques which is used when the outcome variable is categorical. The algorithm predicts a categorical variable by building a training model based on several numerical and categorical variables in the dataset. In this study, the outcome variable is the student course grade. The grade has been encoded as ‘0’ or ‘1’ where ‘0’ denotes scores less than 60 percent and ‘1’ denotes scores above 60 percent. The encoding is done according to the grade scale of the course where scores less than 60 indicate an F grade.

## C. Feature selection

The cleansed data is used for selection of attributes also known as feature selection that determines which predictors should be included in the model to flawlessly contribute to the output variable. It filters out irrelevant and redundant features which can have a negative effect on the training model's performance. After determining the correlation factor shown in Figure 2, Figure 3 and Figure 4, the uncorrelated factors were eliminated to enhance the performance of the training model.

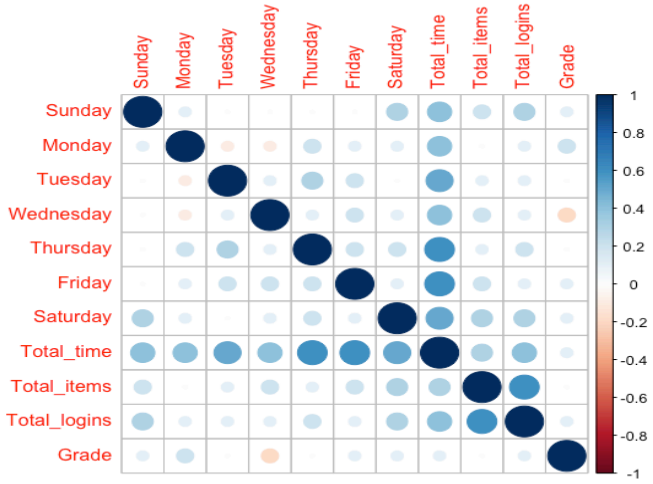


**Figure 2. Correlation plot for days of the week**

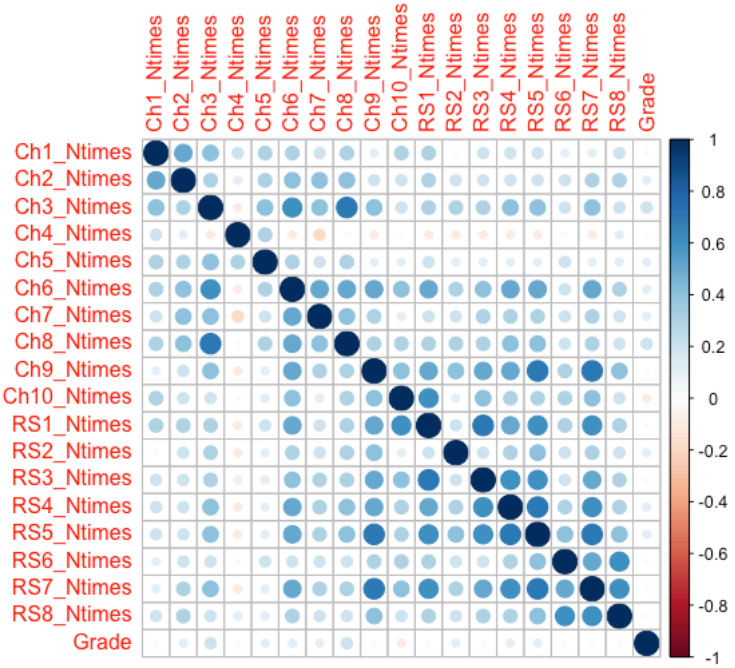


**Figure 3. Correlation plot for chapters**

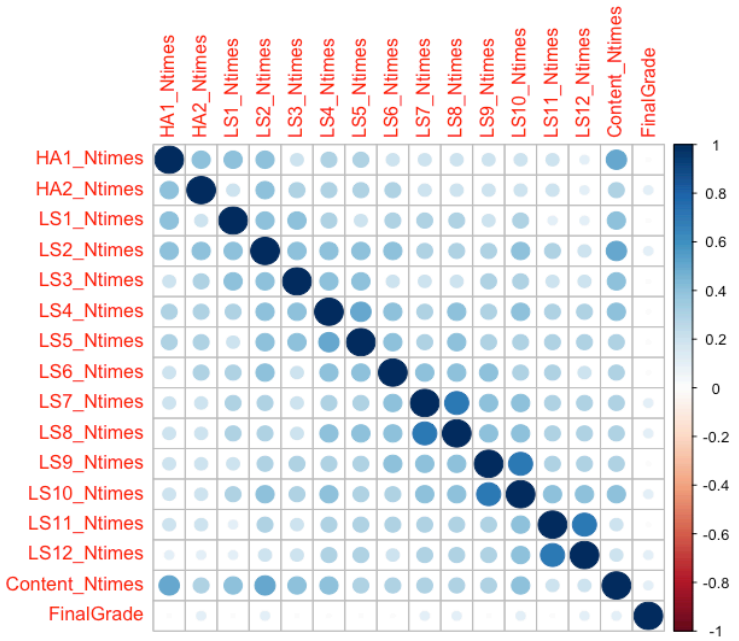


**Figure 4. Correlation plot for homework/lab assignments**

The correlation plots provide information about the relationship between variables. Figure 2 depicts that grade has a positive correlation with Monday, and it can be inferred that if the student invests more time on the course on Monday, they have a better chance of scoring a good grade. This indicates that there is a relation between the due date and the time spent on the course before the due date. All the variables in Figure 3 and Figure 4 represent the number of times each item was accessed by the students. Figure 3 illustrates final grades have a positive correlation with most of the chapters from which it can be concluded that it is essential for students to learn all the chapters in order to improve their final grade. The same can be inferred from Figure 4 about the homework and lab assignments. From the correlation plots, it can be observed that final grade has a positive correlation with most of the attributes except for Wednesday in Figure 2 and Chapter 10 in Figure 3 for which it has a negative relationship. These attributes are not included in the modeling process.

### D. Training model process

The dataset has to be divided into training data and test data for modeling where training data is used to train the model and test data is used to validate the model. We choose to divide the dataset into 90% training data and 10% test data which infers that 90% of the dataset is used to train the model and the remaining 10% of the dataset is used to validate the performance of the model. The following classification models were used to predict the outcome.

1) Support Vector Machine (SVM)
2) Naïve Bayes
3) K-Nearest Neighbors (KNN)
4) Linear Discriminant Analysis (LDA)

Support Vector Machine [1] is a very powerful and flexible modeling technique. It implements classification to categorize data points even in extreme cases via a decision boundary which is referred as a hyperplane. The algorithm categorizes new data based on the optimal hyperplane which is obtained from the training data. Naïve Bayes [2] is based on the Bayes' theorem and is useful for large datasets. It assumes that all the predictors are independent of each other. K-Nearest Neighbors [3]makes use of all the available cases in the data and classifies new data based on the similarity measure which is usually the distance function. Removing noisy and irrelevant data can improve the performance of the model. Linear Discriminant Analysis [4] is a simple classification technique which is based on searching for a linear combination of predictors that categorize the target.

**E. Model evaluation**

Model evaluation has to be done in order to understand how well the model is working. Some of the common metrics used to evaluate the performance of classification model are accuracy, precision, and specificity of the model which can be calculated from the confusion matrix, ROC chart, and AUC curve [17]. Accuracy, specificity and AUC are considered in this study since they are the most common metrics used. Accuracy in Figure 5 is defined as the number of items categorized correctly divided by the total number of items. It illustrates the portion of total number of predictions that are correct. Specificity in Figure 5 is the portion of actual negative cases that are identified correctly. The higher value of AUC [11] indicates a better classifier. A perfect classifier will have an AUC of 1.

| Confusion Matrix | | Target | |
|---|---|---|---|
| | | Positive | Negative |
| Model | Positive | a | b |
| | Negative | c | d |

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} \qquad \text{Specificity} = \frac{d}{b+d}$$

**Figure 5: Classification model metrics**

## 4. Numerical Results

According to the analysis, Support Vector Machine (SVM) works best for our dataset. Table 2 provides the values of evaluation metrics for all the algorithms used in this analysis. The accuracy and specificity were derived from the confusion matrix and the AUC is obtained using its respective function in R Studio. SVM and K-Nearest Neighbor have the same accuracy and specificity, but the AUC is better for SVM. The ideal value of AUC should be between 0.5 and 1. SVM is able to classify 87.5% of the data accurately with 100% specificity and AUC of 0.5.

**Table 2. Metric values pertaining to each algorithm**

| | Accuracy | Specificity | AUC |
|---|---|---|---|
| Support Vector Machine | 87.5 | 100 | 0.5 |
| Naïve Bayes | 75 | 85.71 | 0.429 |
| K-Nearest Neighbor | 87.5 | 100 | 0.429 |
| Linear Discriminant Analysis | 62.5 | 71.43 | 0.357 |

## 5. Conclusion and Future work

Students' grades have a relationship with several factors like the number of hours spent on the course on each day, number of hours spent on particular course content, lab assignments, homework assignments, total number of logins, etc. Using these factors, we built a model which can successfully predict students' performance based on their navigational behavior. According to our analysis, Support Vector Machine is more accurate and effective than the other algorithms. This model can help the instructors in providing better guidance for the students. The data is not normally distributed because the output variable contains records with more ones and fewer zeros where we considered 1 for scores above 60% and 0 for scores below 60%. The data has very few students with scores below 60% and hence it is not normally distributed. The models can perform better if we have a larger dataset.

For our future work, we plan to include several other sections that will allow us to experiment with more data. This would help us in improving the accuracy and other evaluation metrics we considered in our study thereby giving us accurate predictions of the students' performance.

## References

[1] M. Kuhn and K. Johnson, "Nonlinear Classification Models - Support Vector Machines," in *Applied Predictive Modeling*, New York, Springer, 2013, pp. 343-350.

[2] M. Kuhn and K. Johnson, "Nonlinear Classification Models - Naive Bayes," in *Applied Predictive Modeling*, New York, Springer, 2013, pp. 353-358.

[3] M. Kuhn and K. Johnson, "Nonlinear Classification Models - K Nearest Neighbors," in *Applied Predictive Modeling*, New York, Springer, 2013, pp. 350-353.

[4] M. Kuhn and K. Johnson, "Discriminant Analysis and Other Linear Classification Models - Linear Discriminant Analysis," in *Applied Predictive Modeling*, New York, Springer, 2013, pp. 287-297.

[5] J. Superby, J. Vandamme and N. Meskens, "Determination of factors influencing the achievement of the first-year university students using data mining methods," in *In Proceedings of the workshop on educational data mining, ITS'06*, 2006.

[6] J.-L. Hung and K. Zhang, "Revealing Online Learning Behaviors and Activity Patterns and Making Predictions with Data Mining Techniques in Online Teaching," *MERLOT Journal of Online Learning and Teaching,* vol. 4, no. 4, 2008.

[7] F. Widyahastuti and V. U. Tjhin, "Predicting Students Performance in Final Examination using Linear Regression and Multilayer Perceptron," in *In Human System Interactions (HSI), 2017 10th International Conference on(pp. 188- 192). IEEE.* , 2017.

[8] M. Kuhn and K. Johnson, "Linear Regression," in *Applied Predictive Modeling*, New York, Springer, 2013, p. 105.

[9] A. G. Parlos, B. Femandez, A. F. Atiya, J. Muthusami and W. K. Tsai, "An Accelerated Learning Algorithm for Multilayer Perceptron Networks," *IEEE TRANSACTIONS ON NEURAL NETWORK,* vol. 5, no. 3, 1994.

[10] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani and K. L. Addison, "A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes," in *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[11] J. Huang and C. X. Ling, "Using AUC and Accuracy in Evaluating Learning Algorithms," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,* vol. 17, no. 3, 2005.

[12] Y.-H. Hu, C.-L. Lo and S.-P. Shih, "Developing early warning systems to predict students' online learning performance," *Computers in Human Behavior,* vol. 36, p. 469–478, 2014.

[13] L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an ‘‘early warning system" for educators: A proof of concept," *Computers & Education,* vol. 54, pp. 588-599, 2010.

[14] J. W. You, "Identifying significant indicators using LMS data to predict course achievement in online learning," *Internet and Higher Education,* vol. 29, pp. 23-30, 2016.

[15] B. K. Baradwaj and S. Pal, "Mining Educational Data to Analyze Students‟ Performance," *International Journal of Advanced Computer Science and Applications,* vol. 2, no. 6, 2011.

[16] S. v. Buuren, in *Flexible Imputation of Missing Data, Second Edition*, CRC Press Taylor & Francis Group, 2018, p. 12.

[17] M. Kuhn and K. Johnson, "Measuring Performance in Classification Models," in *Applied Predictive Modeling*, New York, Springer, 2013, pp. 254-266.